Erschienen in: Klawitter, Jana/Lobin, Henning/Schmidt, Torben (Hrsg.): Kulturwissenschaften Digital. Neue Forschungsfragen und Methoden. S. 155-176 - Frankfurt am Main: Campus, 2012.

Linguistische Annotationen für die Analyse von Gliederungsstrukturen wissenschaftlicher Texte

Harald Lüngen, Mariana Hebborn

1. Einleitung

Ziel des kulturwissenschaftlichen Projekts »Die Ordnung von Wissen in Texten« ist die linguistische Beschreibung von Gliederungskonstruktionen anhand einer korpusbasierten Analyse von Gliederungen wissenschaftlicher Texte. Ausgehend von der Hypothese, dass die Struktur akademischer Texte die Wissensstruktur, die im Text aufgebaut wird, widerspiegelt, wird untersucht, wie visuell-hierarchische Gliederungssysteme aufgebaut sind und wie Wissensstrukturen in ihnen kodiert sind. Konkret soll ein Prototyp für eine Ontologieinduktion entwickelt werden, in dem Gliederungsstrukturen genutzt werden, um automatisch Wissen über die semantischen Konzepte einer Domäne und die Relationen, die zwischen ihnen gelten, abzuleiten. Als Gliederungen werden insbesondere die durch Kapitel, Abschnitte und Unterabschnitte mit ihren Überschriften und Zwischenüberschriften gegebenen Strukturen verstanden.

[ ] Einführung **Pädagogik**
  [D] Ausgewählte **Subdisziplinen** und **Fachrichtungen**
    [1] **Erlebnispädagogik**
    [2] **Erwachsenenbildung**
    [3] **Gesundheitspädagogik**

Listing 1: Gliederungsfragment mit fünf Überschriften aus Raithel et al., 2007.

Als Beispiel für eine Gliederungskonstruktion, aus der ontologisches Wissen extrahiert werden kann, zeigt Listing 1 ein Gliederungsfragment aus einem Lehrbuch für Erziehungswissenschaft[1] mit fünf Überschriften auf

1 Raithel Jürgen; Dollinger, Bernd; Hörmann, Georg (2007), *Einführung in die Pädagogik. Begriff – Strömungen – Klassiker – Fachrichtungen*, 2. Aufl., Wiesbaden.

drei Einbettungsebenen. Die Überschrift der ersten Ebene enthält den Term (Fachterminus beziehungsweise sprachlicher Ausdruck für ein Domänenkonzept) *Pädagogik*, die Überschrift der zweiten Ebene enthält die domänenneutralen, relationalen Substantive *Subdisziplinen* und *Fachrichtungen* und in den Überschriften der dritten Ebene finden sich die drei Terme *Erlebnispädagogik, Erwachsenenbildung,* und *Gesundheitspädagogik*. Sie bezeichnen jeweils Subkonzepte des übergeordneten Konzepts, das durch den Term *Pädagogik* benannt wird. Die Subkonzept-Relation (auch *Hyponymie*) ist eine zentrale Relation beim Aufbau von Begriffshierarchien oder -netzwerken (*Ontologien*). Selbst wenn wir die Bedeutung des Wortes nicht kennen würden, gingen wir davon aus, dass *Erwachsenenbildung* ein Subkonzept oder Teilbereich der Erziehungswissenschaften ist. Das liegt im Wesentlichen an der Einbettungsstruktur der Überschriften in dem Gliederungsfragment, an den Schlüsselwörtern (hier *Disziplin* beziehungsweise *Richtung)* in der Überschrift der zweiten Ebene und an den jeweiligen Singularmarkierungen der Terme und den Pluralmarkierungen der Schlüsselwörter.

Um Gliederungsfragmente und Gliederungskonstruktionen anhand realistischer Daten zu untersuchen, wurde ein Korpus wissenschaftlicher Lehrbücher in digitaler Form zusammen gestellt und korpustechnologisch aufbereitet. Unsere zentralen Fragestellungen für die Wahl und die projektspezifische Ausprägung der Methoden waren: Wie muss das Korpus konzipiert sein, damit die Forschungsfrage (Ermittlung von Gliederungskonstruktionen) bearbeitet werden kann? Welche linguistischen Informationsebenen sollten im Korpus ausgezeichnet werden, welche Einheiten und Relationen auf diesen Informationsebenen? Fragen, die bei der Umsetzung beantwortet werden müssen, sind: Welche Ressourcen (Daten oder Software) stehen für linguistische Annotationen bereits zur Verfügung? Welche Ressourcen müssen im Rahmen des Projekts erstellt werden? Was sind die Schritte des Korpusaufbaus, wie muss die linguistische Verarbeitungspipeline für eine dynamische Korpuserstellung aussehen?

Unter den korpuslinguistischen Methoden wird der korpusbasierte, quantitativ-qualitative Ansatz von dem korpusgestützten Ansatz unterschieden (vgl. Lemnitzer/Zinsmeister 2006). Charakteristisch für den korpusbasierten, quantitativ-qualitativen Ansatz sind die Ermittlung von Wortstatistiken, N-Gramm-Analysen[2] und schließlich Kookkurrenzanalysen über rohem Text (Text ohne Formatierungen oder Annotationen), um

2 Statistiken über vorkommende Wortfolgen aus zwei, drei, … n Wörtern.

Forschungsfragen wie beispielsweise die Herausarbeitung von Charakteristika einer Sprachvarietät zu untersuchen. Ein solcher Ansatz wird in dem Beitrag von Saage et al. in diesem Band verfolgt.

Hingegen ist es charakteristisch für den korpusgestützten Ansatz, dass zunächst Texte mit linguistischen Annotationen angereichert werden, um daraufhin Belegstellen oder auch Gegenbeispiele für bestimmte, theoretisch oder anhand von Voranalysen postulierte syntaktische oder textlinguistische Konstruktionen zu ermitteln. Ein solcher Ansatz wird in dem hier beschriebenen Projekt verfolgt. Er eignet sich wesentlich besser für die projektspezifischen Fragestellungen aus dem Bereich der Text- und Dokumentanalyse (siehe auch Stede 2007). Außerdem sei darauf hingewiesen, dass aufgrund der morphologischen und syntaktischen Charakteristika des Deutschen (Komposita, Vielfalt der Flexionsformen, Wortstellungsvarianten) eine Basisaufbereitung von Korpustexten (Lemmatisierung, morphosyntaktische Analyse) auch im quantitativen-qualitativen Ansatz wünschenswert ist, will man ähnliche Ausgangsvoraussetzungen wie für eine Untersuchung des Englischen schaffen.

Im vorliegenden Beitrag steht das von uns entwickelte Methodensetting des Korpusaufbaus und der Korpusaufbereitung, das heißt der linguistischen Annotation mit ihren Teilmethoden, Werkzeugen und Verfahren in dem oben beschriebenen Projekt im Fokus. Wir orientieren uns dabei an dem Leitfaden, der in Lemnitzer/Zinsmeister (2006) vorgestellt wird. Demnach werden beim Aufbau eines neuen Korpus die Schritte Akquisition, Klärung rechtlicher Fragen, Kodierung, Metadaten und Linguistische Annotationen durchgeführt (vgl. Lemnitzer/Zinsmeister 2006: 57ff.).

2. Korpusaufbereitung für Gliederungsanalysen

Für die oben skizzierte wissenschaftliche Fragestellung wurde ein Korpus von zunächst 32 deutschsprachigen wissenschaftlichen Lehrbüchern akquiriert, die für die universitäre Lehre gedacht und einer traditionellen wissenschaftlichen Disziplin, wie Linguistik, Pädagogik, Psychologie oder Biologie, zugeordnet sind. Von wissenschaftlichen Lehrbüchern wird angenommen, dass ihre Gliederungen am ehesten das System oder Teile des Systems der Grundbegriffe einer Disziplin widerspiegeln. Die Untersuchung weiterer, spezialisierterer wissenschaftlicher Texttypen, wie Dissertationen

oder wissenschaftliche Zeitschriftenartikel, soll in einer späteren Projektphase vorgenommen werden. Auch sind im derzeitigen Lehrbuchkorpus die Naturwissenschaften noch unterrepräsentiert, da das Programm der meisten kontaktierten Verlage vorwiegend geisteswissenschaftlich orientiert war.[3]

2.1 Korpusakquisition und Extraktion von Text und Struktur

2.1.1 Rechtliche Fragen

Für den Erwerb von wissenschaftlichen Lehrbüchern in digitaler Form wurden wissenschaftliche Verlage angeschrieben mit der Bitte um die Freigabe einer von uns zusammengestellten »Wunschliste« von Lehrbüchern für unsere Forschungszwecke. In allen Fällen mussten wir schriftlich versichern, dass die ausgelieferten Bände ausschließlich von Projektmitarbeitern für die beschriebenen Projekte genutzt werden und keinesfalls an Dritte weiter gegeben werden. Dass annotierte Sprachkorpora nicht der wissenschaftlichen Öffentlichkeit zur Verfügung gestellt werden können, ist zwar grundsätzlich bedauerlich, aber im Fall von vollständigen wissenschaftlichen Lehrbüchern, die fast alle noch im Handel sind, eine verständliche Beschränkung.

2.1.2 Extraktion von Text und Struktur aus PDF-Dateien

Die elektronischen Lehrbücher wurden von den Verlagen zumeist in Form von PDF (*Portable Document Format*)-Dateien ausgeliefert, was eine große Herausforderung darstellte, da der Gesamttext mit seiner Dokumentstruktur (Zwischenüberschriften, Unterkapitel und Einbettung der Kapitel) so extrahiert werden sollte, dass er auf einfache Weise in XML-Markup überführt werden konnte. Da PDF eigentlich eine Beschreibungssprache für Seitenlayout ist, funktioniert dies nur für bestimmte PDF-Dateien. Wir testeten unterschiedliche Software, mit der Text, HTML mit Text oder XML mit Text aus dem PDF extrahiert werden kann. Die besten Ergebnisse wurden mit der ›Export XML‹-Option der kommerziellen Software

3 Wir danken den Verlagen Facultas, Haupt, Narr/Francke/Attempto, Springer, UTB, Vandenhoeck & Ruprecht und Wissenschaftliche Buchgesellschaft, die uns freundlicherweise digitale Versionen von Lehrbüchern zur Verfügung gestellt haben.

Adobe Acrobat Pro (Version 9) erzielt, mit der im Idealfall die Dokumentstruktur in einem XML-Format, das wir nach seinem Wurzelelement als TaggedPDF-doc bezeichnen, erzeugt werden konnte.

```
<TEI>
  <teiHeader>...</teiHeader>
  <text>
    <front>
      <div type="frontstuff">[…]</div>
      <div type="contents">[…]</div>
      […]
    </front>
    <body>
      […]
      <div type="section">
        <head xml:id="head_id_119">
          <num type="struct">5.3.3.4</num>
          <label>Grenzen des Wohlfahrtsstaates</label>
        </head>
        <p>...</p>
        <p>...</p>
        <div type="section">[…]</div>
        <div type="section">[…]</div>
      </div>
    </body>
  <text>
</TEI>
```

Listing 2: Gliederungsstruktur in TEI.[4] Überschrift aus dem Lehrbuch Hermann Adam, 2007[5].

4 Das Element <label> wird hier für die Zeichenkette, die die eigentliche Überschrift (ohne Nummerierung) enthält, verwendet, was nicht ganz seiner Definition in den TEI P5 Guidelines entspricht.

5 Bausteine der Politik. Eine Einführung, 1. Aufl., Wiesbaden 2007.

XML (*Extensible Markup Language*) ist ein offener Standard des *World Wide Web Consortium* (W3C) zur Kodierung von Daten, der in der Linguistik mittlerweile als Standard zur Auszeichnung von Textdaten im Allgemeinen und von Korpusdaten im Besonderen gilt.

Eine in der Linguistik, aber auch in den übrigen Geisteswissenschaften weit verbreitete XML-Anwendung ist TEI, benannt nach und definiert von der *Text Encoding Initiative*.

Aufgrund des geschätzten Bedarfs in unserem Projekt wurde als Dokumentgrammatik ein TEI-Schema für eine Teilmenge von Elementen und Attributen des Gesamtumfangs des TEI-Standards entwickelt, mit denen alle Lehrbücher im Korpus ausgezeichnet wurden. Das Schema enthält TEI-Elemente und Attribute zur Auszeichnung von Metadaten, Abschnitten und Zwischenüberschriften. In Listing 2 sieht man beispielsweise die Verwendung der TEI-Elemente `<div>` zur Markierung der (verschachtelten) Abschnittsstruktur und von `<head>` zur Markierung eines Überschriftbereichs.

Das Vorgehen bestand folglich darin, mittels der Software *Adobe Acrobat Pro* TaggedPDF-Dateien aus den PDF-Dateien zu extrahieren und anschließend durch Konvertierungsroutinen in die TEI-Annotation zu überführen.

Zunächst wurde das gewonnene TaggedPDF-Dokument auf Weiterverarbeitbarkeit überprüft. Zwei Drittel aller akquirierten Lehrbücher mussten an dieser Stelle aussortiert werden, da ihre Dokumentstruktur nicht oder nur in extrem korrumpierter Form extrahiert werden konnte. Das restliche Drittel wurde mittels einer Verkettung von XSLT-Stylesheets und einigen manuellen Korrekturen in das projektspezifische TEI-Format überführt. Besagte manuelle Korrekturen betrafen bei der Extraktion im TaggedPDF-Dokument verbliebene unmarkierte Überschriften, fälschlich als Überschriften markierte Bereiche sowie fehlerhafte Einbettungsstrukturen und waren bei einigen Büchern recht aufwändig. Die Stylesheets mussten für jedes PDF-Dokument an einigen Stellen neu angepasst werden, an denen die TaggedPDF-Dokumente keine einheitliche Struktur bezüglich Überschriften und Einbettungen aufwiesen.

2.2 Metadaten

Die Metadaten beinhalten die bibliografischen Angaben zu einem Lehrbuch sowie Angaben zu ihrer Bearbeitung im Projekt und ihrem Status im Gesamtkorpus. Sie wurden ebenfalls nach dem TEI-Standard, das heißt in einem `<teiheader>`-Bereich, erstellt. Der `<teiheader>`-Bereich wurde als separate Datei angelegt, die durch eine `xi:include`-Referenz in die TEI-Gliederungsstruktur eingebunden ist.

2.3 Textbereinigung und Normalisierung

Extrahierter Text	Bereinigter Text
Wir soUten	Wir sollten
sinnvoU	sinnvoll
libereinstimmen	übereinstimmen
erflillt	erfüllt
engHsche	englische
modem	modern
erlautert	erläutert
re-den	reden
person-Hchste	persönlichste

Tabelle 1: Beispiele für OCR-Phänomene beim Extrahieren von Text aus PDF.

In unserem Projekt betraf dieser Schritt die Überprüfung und – wo erforderlich – Korrektur der Zeichenkodierung der extrahierten Lehrbuchtexte, die Eliminierung von Kopfzeilen aus dem fortlaufenden Text und die Eliminierung von Silbentrennungen, wie in den folgenden Abschnitten genauer dargestellt wird.

2.3.1 Zeichenkodierung

Im Zielzustand sollte das Korpus einheitlich in der Unicode-Kodierung UTF-8 vorliegen, da diese Kodierung mittlerweile als Standard für Korpora angesehen werden kann. So wird die UTF-8-Kodierung zum Beispiel auch

in Lemnitzer/Zinsmeister (2006) für Korpora empfohlen, da mit ihr tausende von Zeichen darstellbar sind, das heißt auch Sonderzeichen, wie zum Beispiel mathematische Symbole und griechische Buchstaben, im Gegensatz zu dem älteren, aber auch gebräuchlichen Standard ISO-8859-1 für westeuropäische Alphabete, der nur maximal 256 darstellbare Zeichen umfasst.

Schwierigkeiten ergaben sich dadurch, dass Adobe Acrobat Pro zwar den Text in UTF-8 exportierte, jedoch für einige Texte nicht die gewünschten Zeichen(folgen) ausgab (zum Beispiel ›a‹ statt ›ä‹, ›U‹ statt ›ll‹ oder ›m‹, statt ›rn‹) oder Ligaturen falsch erkannt wurden, wie es typisch für eine optische Zeichenerkennung (*Optical Character Recognition*, OCR) ist. Außerdem erschienen Worttrennungen wie ›ver wendet‹ statt ›verwendet‹ und Zusammenschreibungen wie ›diesubjektive‹ anstatt ›die subjektive‹. Offenbar wird beim Export aus PDF bei einigen Texten von Adobe Acrobat Pro eine OCR durchgeführt, weitere Beispiele finden sich in Tabelle 1. Für einige der anvisierten Text-Analyseverfahren im Rahmen der Ontologieinduktion (beispielsweise Term-Identifikation, bei der die Häufigkeit von Wörtern und Phrasen im Fließtext ermittelt wird, aber auch für die lexikonbasierten Tagger, siehe Abschnitt 2.4) ist eine einheitliche Schreibung Voraussetzung, daher wurden die häufigsten dieser falschen Zeichenfolgen ermittelt und durch ein Bereinigungsskript, das reguläre Ausdrücke anwendet, abgefangen. Für das Zusammenfügen falscher Getrenntschreibungen wurden beispielsweise Wortformen-Bigramme, von denen mindestens ein Bestandteil keine gültige Wortform war, zusammengezogen, wenn das Resultat ein Wort ergab. Der Wortstatus wurde dabei anhand der sehr großen Leipziger deutschen Wortschatzliste[6] überprüft. Die dargestellten Bereinigungsschritte mussten für fünf der 32 Lehrbücher durchgeführt werden.

Ein weiteres Problem war, dass einige der verwendeten Annotationstools nur die ISO-Kodierung (siehe oben) als Eingabe und/oder Ausgabe akzeptierten (beispielsweise *Machinese Syntax*, siehe Abschnitt 2.4). Zu diesem Zweck mussten in der Bearbeitungspipeline (siehe Abschnitt 2.6) entsprechende Konvertierungen und Rekonvertierungsschritte vorgesehen werden. Deswegen ist im Endzustand zwar das Korpus in UTF-8 kodiert, verwendet werden aber aus diesem Zeichensatz nur diejenigen Zeichen, die auch in ISO-8859-1 vorkommen. Das heißt, dass bestimmte Sonderzei-

6 Vgl. http://wortschatz.uni-leipzig.de, vgl. auch Quasthoff (1998).

chen, falls in sinnvoller Weise möglich, durch Zeichen oder Zeichenfolgen in ISO-8859-1 ersetzt wurden, zum Beispiel ›—‹ (»m-dash«) auf ein einfaches ›-‹ (»minus«). Andere, wenig gebräuchliche Sonderzeichen dagegen wurden notfalls getilgt.

2.3.2 Kopfzeilen und Silbentrennung

Da PDF eine Seitenlayout-Beschreibungssprache ist, werden beim Export der Text- und Dokumentstruktur auch fortlaufend Kopfzeilen ausgegeben, die die Überschrift des aktuellen Kapitels wiedergeben und in vielen Büchern auf jeder Seite erscheinen. Solche Kopfzeilen wurden mittels eines Perl-Skripts nachträglich eliminiert, da sie sich teilweise häufig wiederholen und somit die auf Worthäufigkeiten beruhenden Differenzanalysen bei der Term-Identifikation verzerren würden.

Eine weitere Folge des Exports aus PDF sind Getrenntschreibungen, die von Silbentrennungen am Zeilenende stammen wie ›re- den‹; solche werden ebenfalls durch ein Skript unter Abgleich mit der Leipziger Referenzwortliste zusammengezogen.

Für unsere Forschungsfragen war es am wichtigsten, dass die Überschriftenbereiche der Lehrbücher einwandfrei von den beschriebenen PDF-Export-Fehlern befreit wurden. In den Fließtextbereichen der Lehrbücher (das heißt im überwiegenden Textanteil) wurden nur die auffälligsten beziehungsweise häufigsten und mit einfachen regulären Ausdrücken zu korrigierenden Fehler behoben.

2.4 Relevante Annotationsebenen und verwendete Annotationswerkzeuge

In diesem Abschnitt wird die Aufbereitung durch linguistische Annotationen behandelt, die erforderlich ist, damit eine Ontologieinduktion durchgeführt werden kann. Laut Maedche/Staab (2004) umfasst die Ontologieinduktion die Extraktion von Konzepten (in diesem Projekt angenähert durch Term-Extraktion), und die Extraktion semantischer Relationen zwischen den ermittelten Konzepten. Insbesondere sollen domänenspezifische semantische Relationen über so genannte Gliederungskonstruktionen (Lüngen/Lobin 2010, Lüngen/Hebborn 2010) ermittelt werden – eine Erweiterung der Methode der Extraktion mit Lexiko-Syntaktischen Mustern (*Lexico-Syntactic Pattern Extraction* – LSPE, nach Hearst 1998) für Textglie-

derungen. Neben der Annotation der Dokumentstruktur durch TEI-Markup werden daher dem Korpus durch bestimmte Werkzeuge einige linguistische Annotationsschichten hinzugefügt. Dies erfolgt im Rahmen XML-basierter Multi-Ebenen-Annotationen (Goecke et al. 2010), das heißt, dass der Text für jede Annotationsschicht dupliziert und dann mit Markup versehen wird, so dass mehrere eigenständige XML-Dateien entstehen. Im Folgenden werden die drei Annotationsschichten Tokenisierung/Lemmatisierung/POS-Tagging, Chunking und Morphologie/Syntax beschrieben.

2.4.1 Tokenisierung, Lemmatisierung und POS-Tagging

Die Tokenisierung eines Fließtextes beinhaltet die Markierung der Wörter und gegebenenfalls Mehrworteinheiten wie *ad hoc* als einzelne Tokens. Die Idee ist, dass sich alle höherrangigen Annotationen im Korpus (zum Beispiel syntaktische Struktur oder Term-Markierung) auf die entstandenen Tokens beziehen, so dass zum Beispiel annotierte Terme an Token-Grenzen beginnen und aufhören.

Lemmatisierung betrifft die Annotation von Grundformen für jedes Wort (beispielsweise »empirisch« für »empirische«), und im POS-Tagging werden Wortarten als Part-of-Speech-Tags wie N, V, A ausgezeichnet. Hier ist es von Interesse, dass sich das Inventar der möglichen POS-Tags nach einem definierten Standard richtet. Das Werkzeug TreeTagger (Schmid 1994) verwendet das so genannte Stuttgart-Tübingen-Tagset (STTS, Schiller et al. 1999) das als Quasi-Standard für das Deutsche angesehen werden kann. Für die Tokenisierung, Lemmatisierung und das POS-Tagging wurde es, da es zudem weit verbreitet und kostenlos ist, im Projekt eingesetzt. Das Spaltenformat der TreeTagger-Ausgabe wurde durch ein Perl-Skript in eine XML-Annotation konvertiert; die Beispielannotation der Überschrift »Die empirische Ermittlung von Schichten« findet sich in Listing 3.

2.4.2 Chunking

Mit dem TreeTagger wird ein Chunker-Modul ausgeliefert, welches phrasale Einheiten in Form nominaler, präpositionaler oder verbaler Chunks (NC, PC und VC) markiert. Der Chunker basiert auf den POS-Tags des TreeTaggers. Da der Chunker keinen Input verarbeiten kann, welcher Lemma-Annotationen beinhaltet, wurden TreeTagger mit Lemmatisierung

und Chunker getrennt angewendet und die Ergebnisse mit Hilfe eines Skripts rekombiniert. Ein in XML konvertierter Output der TreeTagger-Anwendung inklusive Chunking ist in Listing 3 zu sehen, es enthält unter anderem den nominalen Chunk »die empirische Ermittlung« und der präpositionale Chunk »von Schichten«. Die `<token>`-Elemente markieren die Tokenisierung, im Attribut `lemma` ist die Grundform (das Lemma) der Wortform angegeben, dabei wird `"d"` als die Grundform des bestimmten Artikels angegeben und `"@card"` als die Grundform einer Nummerierung.

```
<NC>
   <token pos="CARD" lemma="@card@">8.5</token>
</NC>
<NC>
   <token pos="ART" lemma="d">Die</token>
   <token pos="ADJA" lemma="empirisch">
      empirische</token>
   <token pos="NN" lemma="Ermittlung">
      Ermittlung</token>
</NC>
<PC>
   <token pos="APPR" lemma="von">von</token>
   <token pos="NN" lemma="Schicht">Schichten</token>
</PC>
```

Listing 3: POS-Tagging und Chunking (TreeTagger). Überschrift aus dem Lehrbuch Heinz Abels, 2007[7].

Die Angaben der Wortart im Attribut `"pos"` stehen für folgende Kategorien (vgl. Schiller et al. 1999):

`CARD`: Kardinalzahl
`ART`: Artikel
`ADJA`: attributives Adjektiv
`NN`: Nomen appellativum (Gattungsname, das heißt normales Substantiv)
`APPR`: Adposition: Präposition

7 *Einführung in die Soziologie. Band 1: Der Blick auf die Gesellschaft*, 3. Aufl., Wiesbaden 2007.

2.4.3 Morphologie und Syntax

Für die Untersuchung von Gliederungsstrukturen ist es oft relevant, in welcher morphosyntaktischen Form eine Wortform in einer Überschrift erscheint, ob beispielsweise im Numerus Plural und im Kasus Genitiv. Ebenfalls ist häufig relevant, ob eine Wortform, zum Beispiel ein Term, als syntaktischer Kopf einer Überschrift fungiert oder als eingebettete Phrase erscheint.

```
<token id="w697">
    <text>Die</text>
    <lemma>die</lemma>
    <depend head="w699">det</depend>
    <tags>
      <morpho>DET Def FEM SG NOM</morpho>[…]
    </tags>
</token>
<token id="w698">
    <text>empirische</text>
    <lemma>empirisch</lemma>
    <depend head="w699">attr</depend>
    <tags><morpho>A FEM SG NOM</morpho></tags>
</token>
<token id="w699">
    <text>Ermittlung</text>
    <lemma>ermittlung</lemma>
    <depend head="w696">main</depend>
    <tags><morpho>N FEM SG NOM</morpho>[…]</tags>
</token>
```

Listing 4: Dependenzstruktur in XML nach Machinese Syntax.

Um solche Merkmale zu taggen, bedarf es einer syntaktischen Analyse des gesamten Kontexts einer Wortform. Eine solche Syntaxanalyse für deutsche Texte bietet der kommerzielle Dependenzparser *Machinese Syntax* der Firma Connexor Oy, mit dem alle Texte des Lehrbuchkorpus getaggt wurden. Listing 4 zeigt die syntaktische Analyse für die Beispielüberschrift. »`FEM SG NOM`« sind beispielsweise die morphosyntaktischen Merkmale der Wortform *empirische* (Tag `<morpho>`) in der oben angegebenen Überschrift. Die ID-Referenz (IDREF) `head="w1124"` verweist auf das id-Attribut

des folgenden Tokens und gibt an, dass dieses der syntaktische Kopf zu dem aktuellen Token darstellt. Die Gesamtheit der IDREF-Verweise in den head-Attributen eines Satzes oder einer Phrase spannen einen (im Idealfall korrekten) syntaktischen Dependenzbaum nach der *Functional Dependency Grammar* (FDG, vgl. Tapanainen/Järvinen 1997) auf (vgl. Abbildung 1).

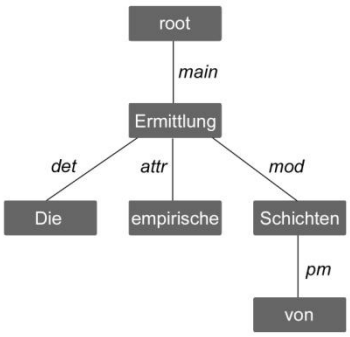

Abbildung 1: Visualisierung der Dependenzstruktur nach Machinese Syntax.

Der Dependenzparser liefert auch eine eigene Tokenisierung und Lemmatisierung sowie ein POS-Tagging, dieses allerdings nach einem proprietären Tag-Set. Im Rahmen der XML-basierten Multi-Layer-Annotation stellt es kein Problem dar, auch konkurrierende Annotationsschichten im Korpus zu verwalten. Die Multi-Layer-Annotation erlaubt auch, zu jeder Annotationsschicht eine eigene XML-Dokumentgrammatik zu definieren und/oder eine bereits vorhandene zu verwenden.

2.4.4 Weitere Annotationsschichten

Weitere Annotationsschichten können jederzeit hinzu gefügt werden. Im Projekt gibt es für einen Teil des Lehrbuchkorpus eine weitere Annotationsschicht, in der domänenrelevante Terme im Bereich der Überschriften ausgezeichnet sind. Diese Term-Information dient dem Entwickeln und Testen einer Term-Identifikationskomponente. Ebenfalls für einen Teil des Korpus wurde eine Annotation funktionaler Typen von Überschriften angefertigt, die der Entwicklung einer Klassifikationskomponente für funktionale Typen von Überschriften dient. In der Ontologieinduktion dient die funktionale Markierung von Überschriften als Filter, indem Überschriften

bestimmter Typen keine domänenrelevante Information enthalten, sondern sich zum Beispiel auf ein Textstrukturmodell wissenschaftlicher Arbeiten beziehen, wie die Überschrift »Weiterführende Literatur«.

Für die Annotationsschichten *Terminologie* und *Funktionale Typen* wurden eigene Dokumentgrammatiken in XML Schema und Annotations-Guidelines für die manuelle Annotation mit dem XML-Editor Oxygen[8] entworfen.

2.5 Multi-Ebenen-Annotationen

Die XML-basierte Multi-Ebenen-Annotation erfordert, dass in der Annotationsphase für jede neue Annotationsschicht der Ursprungstext (Primärtext) jeweils kopiert und dann annotiert wird. Es entsteht also für jede Annotationsschicht ein neues, eigenständiges XML-Dokument. Der Vorteil ist, dass auf diese Weise für jede Annotationsschicht eine eigene Dokumentgrammatik (zum Beispiel eine DTD oder ein XML Schema) angegeben werden kann und dass die Annotationsschichten unabhängig voneinander hinzugefügt werden können und somit auch spezialisierte Werkzeuge, wie der TreeTagger und der Machinese Syntax-Parser, für die Annotation eingesetzt werden können. Außerdem können Annotationen mit Spezial-XML-Editoren wie Oxygen erstellt werden, da Textdokumente mit nur einer Annotationsschicht für den Betrachter leicht erfassbar sind (vgl. Goecke et al. 2010). Wichtig bei diesem Vorgehen ist, dass der kopierte Primärtext beim Annotieren nicht verändert wird. Das hat nämlich zur Folge, dass alle Annotationsschichten eines Texts immer aufeinander beziehbar sind eben anhand des identischen Primärtextes. Für Analysen in Form von Anfragen (*Queries*) an ein Lehrbuchdokument mit seinen Annotationen ist ein solcher Bezug der Annotationsschichten untereinander Voraussetzung. Beispielsweise könnte eine Analyseanfrage im Rahmen der Ontologieinduktion lauten: »Zeige mir alle Nominalphrasen im Genitiv an, die sich im Bereich von Überschriften befinden und die als Terme der Domäne gelten.«

8 Vgl. http://www.oxygenxml.com.

```
<xsf:corpusData >
  <xsf:primaryData start="0" end="3564">
    <xsf:primaryDataRef
       uri="abels_soziologie_2007-tei.nxslt.normalized-pd.txt"/>
    </xsf:primaryData>
    <xsf:segmentation>
      …
      <xsf:segment xml:id="seg701" start="2876" end="2889"/>
      <xsf:segment xml:id="seg702" start="2876" end="2879"/>
      <xsf:segment xml:id="seg703" start="2880" end="2889"/>
      …
    </xsf:segmentation>
    <xsf:annotation>
      <xsf:level
         xml:id="abels_soziologie_2007-cnxde.nxslt.normalized-level1">…
        <analysis xmlns="http://www.text-technology.de/cnxde"
           xsf:segment="seg1">
          …
          <token xsf:segment="seg702" id="w684" lemma="von"
             dependHead="w685" dependValue="pm" syntax="@PREMARK"
             morpho="PREP"/>
          <token xsf:segment="seg703" id="w685" lemma="schichten"
             dependHead="w683" dependValue="mod" syntax="@NH"
             morpho="N NEU SG DAT"/>
          …
      </analysis> …
    </xsf:level>
    <xsf:level
       xml:id="abels_soziologie_2007-ttj.nxslt.normalized-level1">
      …
      <ttj xmlns="http://taurus.zmi.uni-giessen.de/ttj"
         xsf:segment="seg1">
         …
         <PC xsf:segment="seg701">
           <token xsf:segment="seg702" pos="APPR" lemma="von"/>
           <token xsf:segment="seg703" pos="NN" lemma="Schicht"/>
         </PC>
         …
      </ttj> …
    </xsf:level>
  </xsf:annotation>
</xsf:corpusData>
```

Listing 5: Annotationsschichten in XStandoff (Ausschnitt). XML-Format nach Stührenberg/Goecke 2008.

Die Information, ob ein Textbereich eine Nominalphrase im Genitiv ist, kann aus der Annotationsschicht der Dependenzstruktur nach Connexor Machinese Syntax (CNX) entnommen werden, die Information, welcher Textbereich eine Überschrift darstellt, befindet sich in der TEI-Annotation

der Textstruktur (TEI) und die Information, welcher Textbereich einen Term der Domäne (Fachbegriff) darstellt, befindet sich in einer weiteren Schicht für Term-Annotationen, die in Abschnitt 2.4 angedeutet wurde. Um derartige Anfragen mit XML-Standards wie XPath und XQuery formulieren zu können, ist also eine integrierte Sicht des Textes eines Dokuments mit allen seinen Annotationsschichten erforderlich. Eine solche integrierte Sicht im formalen Rahmen von XML bietet das Format XStandoff, welches in Stührenberg/Goecke (2008) und Stührenberg/Jettka (2009) vorgestellt wurde. Das zugehörige, frei verfügbare XStandoff-Toolkit[9] beinhaltet Stylesheets, mit denen ein Dokument und seine zugehörigen verteilten Annotationsschichten mit identischem Primärtext ein XML-Dokument gemäß XStandoff konvertiert werden. In Listing 5 ist ein solchermaßen generiertes XStandoff-Dokument nur für den Textbereich »von Schichten«, mit zwei Annotationsschichten, nämlich der Dependenzstruktur nach Connexor Machinese Syntax und der TreeTagger-mit-Chunker-Information (TTJ). Der eigentliche Text befindet sich in einer separaten Datei, auf ihn wird im Element `<xsf:primaryData>` referiert. Unter `<xsf:segmentation>` sind alle Textbereiche, die als Elemente in den diversen Annotationsschichten markiert sind, als `<xsf:segment>` mit ihrem zugehörigen Zeichenbereich im Primärtext in den Start- und End-Attributen aufgeführt. Die eigentlichen Annotationen (im Beispiel zwei Mal jeweils unter `<xsf:level>`) enthalten keinen Primärtext mehr, sondern verweisen durch das `xsf:segment`-Attribut nur noch auf das betroffene Segment unter `<xsf:segmentation>`.

Dadurch können nun über mehrere Annotationsschichten hinweg Anfragen formuliert werden und zwar im Gegensatz zu anderen linguistischen Korpussystemen, unter Verwendung der allgemeinen W3C-Standards und XML-basierten Abfragesprachen XPath, XSLT und XQuery (vgl. Stührenberg/Goecke 2008). Für Beispiele dafür, wie man entsprechende Anfragen in XQuery formuliert, verweisen wir auf Nachfolgepublikationen aus unserem Projekt.

9 Vgl. http://www.xstandoff.net/tk.html.

2.6 Phasen der Korpusaufbereitung – Bearbeitungskette

Die bisher beschriebenen Phasen der Korpusbearbeitung sind in Abbildung 2 in einer schematischen Darstellung der Korpusbearbeitungspipeline zusammen gefasst. Zu Beginn (oben) steht ein Lehrbuch als PDF-Dokument, aus dem ein XML-Dokument exportiert wird.

Das XML-Dokument wurde durch Konvertierungs- (»convert«) und Normalisierungsschritte (»normalise«) in das TEI-Format überführt, welches die Dokumentstruktur abbildet (Abschnitte 2.1 und 2.3). Die Tagger-basierten Annotationsschichten TTR (TreeTagger, siehe Abschnitt 2.4.1), TTC (TreeTagger+Chunker, siehe Abschnitt 2.4.2) und CNX (Dependenzstruktur nach Connexor Machinese Syntax, siehe Abschnitt 2.4.3) werden in drei parallelen Pfaden jeweils einem Duplikat des Primärtextes hinzugefügt.

Die Ausgaben des TreeTagger und TreeTagger+Chunker (TTR und TTC) werden zu einem XML-Dokument namens TTJ zusammengeführt (»merge«). Schließlich werden TTJ, die Ausgabe des Machinese Syntax Parsers (CNX) und die Repräsentation der Dokumentstruktur (TEI) über ichren identischen Primärtext mittels der XStandoff-Stylesheets (Stührenberg/Jettka 2009) kombiniert und in die XStandoff-Repräsentation (XSF) überführt (siehe Abschnitt 2.5).

Aufgrund der Anforderungen der Input- beziehungsweise Output-Formate der verwendeten Tagger erfolgen zwischen den verschiedenen Formaten die erforderlichen Konvertierungen der Zeichenkodierungen zwischen UTF-8 (U8) und ISO-8859-1 (L1). Dass die Konvertierungen so zu erfolgen hatten, wie in Abbildung 2 dargestellt, musste zum größten Teil durch Testläufe mit der jeweiligen Software ermittelt werden.

Die Hintereinanderschaltungen der Programmaufrufe in dem umrandeten Bereich in Abbildung 2 wurden als ein UNIX Shell Script realisiert (alternativ hierzu wäre ein UNIX Makefile auch sehr geeignet), das immer von Neuem aufgerufen wird, wenn neue digitale Bücher aus PDF extrahiert wurden und im TEI-Format vorliegen. Die Metadaten zu den Korpusdateien eines Lehrbuchs werden in einer separaten teiheader-Datei festgehalten, die durch eine `xi:include`-Referenz in die TEI-Datei eingebunden ist (hier nicht dargestellt).

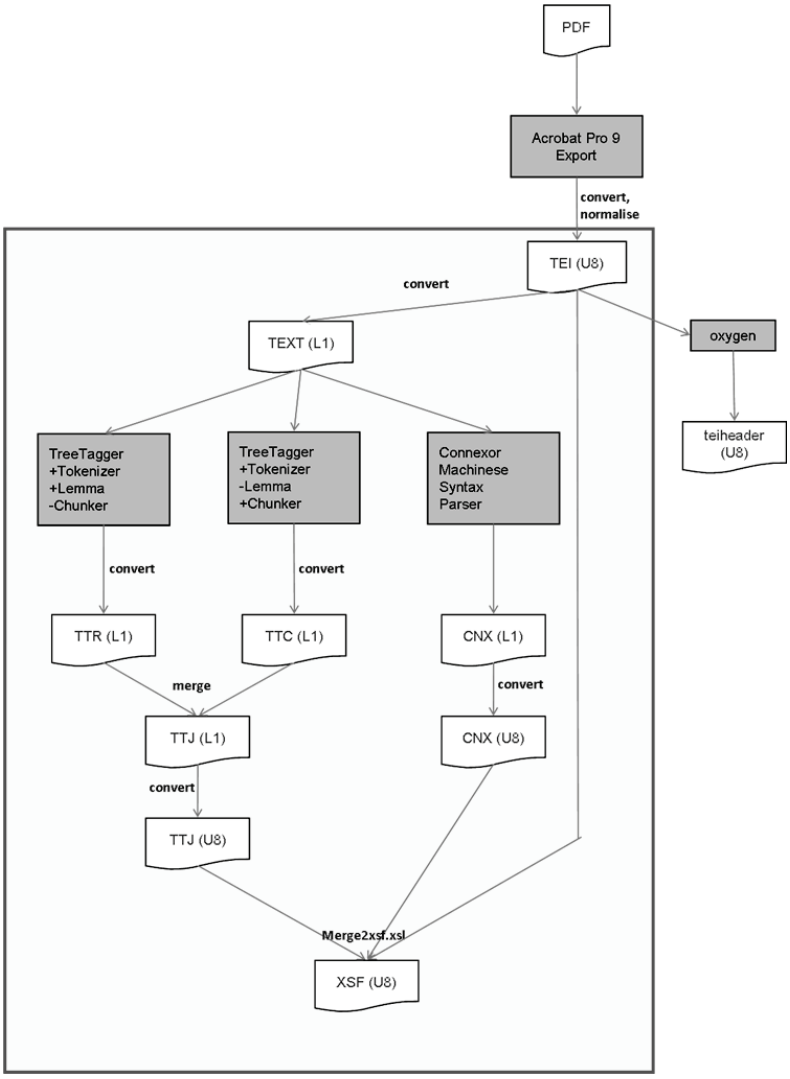

Abbildung 2: Korpusbearbeitungspipeline, Abkürzungen: PDF – Portable Document Format; TEI – Text Encoding Initiative (XML); TEXT – roher Text; TTR – TreeTagger Ausgabe; TTC – TreeTagger-Chunker Ausgabe; TTJ – TreeTagger und TreeTagger-Chunker Kombination (XML); CNX – Connexor Machinese Syntax Ausgabe (XML); XSF – XStandoff Multi-Ebenen-Annotation (XML, Stührenberg/Jettka 2009); L1 – Zeichenkodierung in ISO-8859-1; U8 – Zeichenkodierung in UTF-8 (Unicode).

3. Fazit

In diesem Beitrag wurden Methoden der Korpusaufbereitung für eine korpusgestützte Textanalyse im Rahmen eines Projekts, das sich mit Ontologieinduktion aus Gliederungsstrukturen wissenschaftlicher Texte befasst, dargestellt. Die Arbeitsschritte der Tokenisierung, Lemmatisierung, des POS-Taggings, Chunkings und der Morphologie- und Syntaxannotation im Rahmen einer Multi-Ebenen-Annotation wurden für den Aufbau und die Aufbereitung eines Korpus von wissenschaftlichen Lehrbüchern durchgespielt. Korpusanfragen und Gliederungsanalysen, die mit der entstanden Ressource durchgeführt wurden, waren nicht Teil des vorliegenden Beitrags.

Korpora und Korpusanalysen spielen seit etlichen Jahren eine wichtige Rolle in der Sprachwissenschaft, wo Korpora der empirischen linguistischen Forschung als Datengrundlage dienen, und in der Computerlinguistik, wo Korpora als Trainingsdaten für statistische Modelle und maschinelle Lernverfahren eingesetzt werden. In jüngster Zeit sind sprachliche Ressourcen aber auch in den Blick der übrigen Geisteswissenschaften gelangt. Ein prominentes Beispiel ist das riesige Textarchiv *Google Books* und die daraus extrahierte und im Web verfügbare Ressource der *Google Books Ngrams*. Dazu wurde in einem Artikel von Michel et al. (2011) der Ansatz der sogenannten *culturomics* vorgestellt, in welchem durch quantitative Analysen von Wortvorkommen und Kookkurrenzen in zeitlich gestreuten Korpora historischen und kulturellen Trends nachgespürt wird.

Aber auch die Möglichkeit, wie in dem in diesem Beitrag vorgestellten Projekt, eigene Korpora und Subkorpora für spezifische Fragestellungen und Auswertungen zusammenzustellen und zu annotieren, soll den Geisteswissenschaften zur Verfügung gestellt werden, in Europa beispielsweise durch das EU-geförderte Verbundprojekt CLARIN (*Common Language Resources and Technology Infrastructure*, Váradi et al. 2008). Die Idee ist, dass ein Wissenschaftler, der ein Textarchiv oder Korpus zusammenstellen und aufbereiten möchte, sich gar nicht um technische Fragen, wie Zeichenkodierung, Schnittstellen zwischen Tagging-Werkzeugen und die Implementierung der Verarbeitungspipeline, kümmern muss. In CLARIN wird dafür eine service-basierte Forschungsinfrastruktur entwickelt, mit Hilfe derer in einem Web-Portal die Komponenten einer gewünschten Bearbeitungspipeline angewählt und zu einer so genannten *Tool Chain* zusammen gestellt werden können. Anschließend kann damit ein eigenes Korpus automatisch

aufbereitet werden. (siehe Hinrichs et al. 2010). CLARIN-D, vormals D-SPIN (Deutsche Sprachressourcen-Infrastruktur), ist der vom Bundesministerium für Bildung und Forschung (BMBF) geförderte deutsche Zweig von CLARIN. Das Projekt »Kompetenzzentrum Kulturwissenschaftliche Informationsverarbeitung« unseres LOEWE-Schwerpunkts »Kulturtechniken und ihre Medialisierung« hat von Anfang eng mit D-SPIN kooperiert; im Jahr 2010 wurde beispielsweise gemeinsam die D-SPIN »Sommerschule Sprachressourcen für die Geisteswissenschaften« in Bad Homburg durchgeführt[10]. Solange die Vision der CLARIN-Forschungsinfrastruktur noch nicht abschließend umgesetzt ist (CLARIN ist ein laufendes Projekt), wird für Projekte, in denen Sprachressourcen aufgebaut und aufbereitet werden, aber nach wie vor eine Kooperation mit oder eine Beratung durch eine Einrichtung für kulturwissenschaftliche Informationsverarbeitung oder Digital Humanities empfohlen.

Literatur

Balisage Series on Markup Technologies (Hg.) (2008), *Proceedings of Balisage: The Markup Conference 2008*, Vol. 1, Online-Reihe.

Balisage Series on Markup Technologies (Hg.) (2009), *Proceedings of Balisage: The Markup Conference 2009*, Vol. 3, Online-Reihe.

European Language Resources Association (ELRA) (Hg.) (2008), *Proceedings of LREC 2008*, Marrakesch.

European Language Resources Association (ELRA) (Hg.) (2010), *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valetta/Malta.

Fellbaum, Christiane (Hg.), *WordNet. An Electronic Lexical Database*, Cambridge.

Goecke, Daniela; Metzing, Dieter; Lüngen, Harald; Stührenberg, Maik; Witt, Andreas (2010), »Different views on Markup. Distinguishing Levels and Layers«, in: Andreas Witt; Dieter Metzing (Hg.), *Linguistic Modeling of Information and Markup Languages,* Berlin, S. 1–21.

Hinrichs, Marie; Zastrow, Thomas; Hinrichs, Erhard (2010), »WebLicht: Web-based LRT Services in a Distributed eScience Infrastructure«, in: European Language Resources Association (ELRA) (Hg.), *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valetta/Malta, S. 489–493.

10 Vgl. http://www.dspin-sommerschule.de.

Hearst, Marti A. (1998), »Automated Discovery of WordNet Relations«, in: Christiane Fellbaum (Hg.), *WordNet. An Electronic Lexical Database,* Cambridge, S. 131–153.

Heyer, Gerhard; Wolff, Christian (Hg.) (1998), *Linguistik und neue Medien*, Wiesbaden.

Lemnitzer, Lothar; Zinsmeister, Heike (2006), *Korpuslinguistik. Eine Einführung*, Tübingen.

Lüngen, Harald; Lobin, Henning (2010), »Extracting domain knowledge from tables of contents«, in: Centre for Computing in the Humanities, King's College London (Hg.), *Digital Humanities 2010 Book of abstracts*, London, S. 331–334.

Lüngen, Harald; Hebborn, Mariana (2010), »Konstruktionsgrammatische Analyse von Gliederungsstrukturen«, in: Manfred Pinkal; Ines Rehbein, Sabine Schulte im Walde; Angelika Storrer (Hg.), *Semantic Approaches in Natural Language Processing: Proceeedings of the Conference on Natural Language Processing (KONVENS 2010),* Saarbrücken, S. 151–154.

Maedche, Alexander; Staab, Steffen (2004), »Ontology Learning«, in: Steffen Staab; Rudi Studer (Hg.), *Handbook on Ontologies*, Berlin, S. 173–191.

Michel, Jean-Baptiste B.; Pinker, Steven; Shen, Yuan Kui; Aiden, Aviva P.; Veres, Adrian; Gray, Matthwe K.; The Google Books Team; Pickett, Joseph P.; Hoiberg, Dale; Clancy, Dan; Norvig, Peter; Orwant, John; Nowak, Martin A.; Lieberman-Aiden, Eres (2011), »Quantitative analysis of culture using millions of digitized books«, in: *Science*, 311 (6014), New York, S. 176–182.

Pinkal, Manfred; Rehbein, Ines; Schulte im Walde, Sabine; Storrer, Angelika (Hg.) (2010), *Semantic Approaches in Natural Language Processing: Proceeedings of the Conference on Natural Language Processing (KONVENS 2010),* Saarbrücken.

Quasthoff, Uwe (1998), »Projekt Deutscher Wortschatz«, in: Gerhard Heyer; Christian Wolff (Hg.), *Linguistik und neue Medien*, Wiesbaden, S. 93–99.

Schiller, Anne; Teufel, Simone; Stöckert, Christine; Thielen, Christine (1999), *Guidelines für das Tagging deutscher Textcorpora mit STTS*. Technischer Bericht, Institut für maschinelle Sprachverarbeitung, Universität Stuttgart.

Schmid, Helmut (1994), »Probabilistic Part-of-Speech Tagging using Decision Trees«, in: *Proceedings of the International Conference on New Methods in Language Processing*, S. 154–164.

Staab, Steffen; Studer, Rudi (Hg.) (2004), *Handbook on Ontologies*, Berlin.

Stede, Manfred (2007), *Korpusgestützte Textanalyse. Grundzüge der Ebenen-orientierten Textlinguistik*, Tübingen.

Tapanainen, Pasi; Järvinen, Timo (1997), »A non-projective dependency parser«, in: *Proceedings of the 5th Conference on Applied Natural Language Processing*, Washington, S. 64–71.

Stührenberg, Maik; Goecke, Daniela (2008), »SGF – An integrated model for multiple annotations and its application in a linguistic domain«, in: Balisage Series on Markup Technologies (Hg.), *Proceedings of Balisage: The Markup Conference*